Towards Unified Dynamic Face Landmark Detection

Sebastian Regalado[*,1,2,†] **Varshanth Rao**[*,2,†] **Ruowei Jiang**[2,†] **Parham Aarabi**[1] **Igor Gilitschenski**[1]
[1]University of Toronto [2]ModiFace

Abstract

Although advancements in face landmark detection (FLD) methods continue to push performance boundaries, they overlook two major functional limitations: (1) different network parameters need to be trained independently for each "N-point" benchmark dataset, and (2) a model trained on an "N-point" dataset reliably outputs only the N landmarks. In our work, we first conceptualize Face Part-Anchored Landmark Positions (FPALPs), wherein each landmark is treated as a progression value between zero (start) and one (end) along a face part's contour. Every landmark can be expressed in the FPALP format, irrespective of its source dataset, hence unlocking the ability to unify all "N-point" datasets into a single dataset. Secondly, we represent each landmark with an FPALP-based query, refine it progressively with a cross-modality decoder, and predict its coordinates based on the final representation. Our approach, called Unified Dynamic FLD, embodies these two design choices and streamlines the landmark detection pipeline by enabling (1) a single model to learn on any number of "N-point" datasets, and (2) yield any number of specific landmark predictions by loading the designated landmark queries at runtime. Extensive experiments on multiple benchmark datasets show that our method delivers these benefits while remaining competitive with, and in several cases outperforming existing state-of-the-art methods.

1 Introduction

Face landmark detection (FLD) aims to predict the coordinates of predefined landmarks in facial images. Facial landmarks provide rich and diverse visual cues regarding face shape, face-part positions, and pose information. These are essential for many downstream tasks such as 3D face reconstruction [37, 4], face recognition [15, 33], face expression recognition [23, 25], and more recently facial beauty predictions [2, 12] and face make-up try on [22, 16, 19, 32].

Despite being at the core of numerous applications, FLD algorithms suffer from major inherent drawbacks at both the training and prediction stages due to a *rigid adherence to the landmark layout defined by the training dataset*. Facial images are annotated with different landmark definitions across datasets such as AFLW [48] (19/21 points), 300W [31] (68 points), and WFLW [39] (98 points). Generalizing, we denote an FLD dataset that defines a unique face layout of N landmarks as the term "N-point" dataset. Prior works [40, 46, 10, 18] have advanced FLD performance on these datasets by training on them *individually* using *separate* backbones and/or regression heads, and designing the networks to *output only the dataset-specific N points*. We denote the above as the separate model and common backbone paradigms (see Figure 1) and investigate their demerits in detail below.

Theoretically, each "N-point" dataset can specify facial landmarks according to mutually exclusive semantic definitions. Here, the specialist nature of the separate model paradigm may outweigh the benefits of a model that was trained on multiple datasets through the common backbone paradigm, as only low-level features might be shared. In reality, we observe that *this assumption does not always hold true*. As an example, in Figure 2a, we overlay the landmark predictions output by SLPT [40], a

[*]Equal contribution. [†]Work completed while employed at ModiFace.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

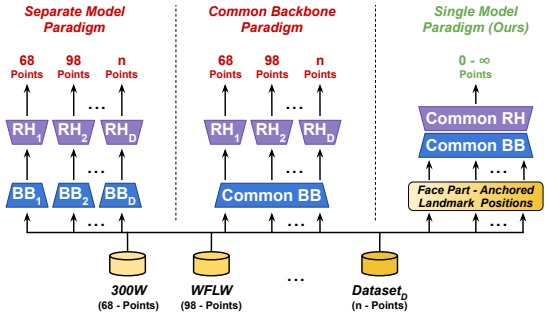
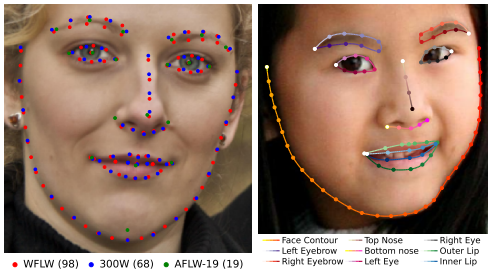


Figure 1: A comparison of the end-to-end training pipeline of prior works' separate model and common backbone paradigms to the single model paradigm implemented by our Unified Dynamic Face Landmark Detection method. BB, RH, and D denote backbone, regression head, and number of datasets, respectively. Based on Face Part-Anchored Landmark Positions, our network can train on the combination of multiple "N-point" datasets and execute an unlimited number of landmark predictions.



(a) (b)

Figure 2: (a) An overlay of the facial landmarks in the AFLW (19-point), 300W (68-point), and WFLW (98-point) formats. The landmark definitions across different datasets are observed to be non-mutually exclusive and strongly semantically related via face parts. (b) Landmarks (excluding pupils) of the WFLW format expressed as Face Part-Anchored Landmark Positions. Each gradient curve transitioning from white to a darker colour indicates the progression from the start to the end of a face part boundary.

state-of-the-art FLD method, that was trained separately on three benchmark datasets; AFLW19 [48], 300W [31], and WFLW [39]. We make two critical observations: (1) facial landmark annotations are semantically anchored to face parts such as eyes, lips, nose, etc., and (2) are often defined to be evenly spaced along a face part boundary [39, 42]. These cause the landmarks in the different "N-point" datasets to be *non-mutually exclusive and strongly semantically related*. Based on these observations, we conceptualize Face Part-Anchored Landmark Positions (FPALPs), in which each facial landmark is first associated with one or more distinct face parts and then assigned a value between 0 and 1 designating a progression point between the start and end of the face part boundary, respectively. We illustrate FPALPs in Figure 2b, wherein most facial landmarks of the WFLW [39] format are anchored to 9 distinct face parts. By indexing facial landmarks as FPALPs calculated on the union of all landmark definitions across the different benchmark datasets, we enable *unified FLD: the ability of an FLD model to be trained end-to-end on the combination of all the considered datasets.*

As noted earlier, during inference, FLD methods trained on an "N-point" dataset outputs only N facial landmarks. Such output rigidity is non-optimal for downstream applications like face direction estimation [1, 34] and FLD stabilization in videos [14, 38] that may utilize only a few sparse facial landmarks, and restrictive for applications like face image animation [44] that require a higher density of accurate facial landmarks. Although higher facial landmark density can be naïvely achieved using interpolation methods, the output accuracy is dependent on a higher N since face parts have non-linear shape. To this end, we construct facial landmark queries on demand using the combination of their FPALPs and the text embedding of the containing face parts, and feed them to a cross-modality decoder-regressor to enable *dynamic FLD: the ability of an FLD model to output the predictions of only the queried landmarks.*

Revisiting Figure 1, our Unified Dynamic FLD, which is founded on the concept of FPALPs, executes a *single model paradigm* that can be trained on the combination of diverse "N-point" datasets, and can

Table 1: An efficiency comparison of different face landmark detection paradigms. D denotes the number of unique "N-point" datasets. $\mathcal{B}$, $\mathcal{H}$, and E_Q denotes the backbone, regression heads, and landmark query encoder respectively.

FLD Paradigm	Efficiency			
	Training Cycles	Inference Calculation	Storage Parameter	Landmark Throughput
Separate Model	D	$D\mathcal{B} + D\mathcal{H}$	$D\mathcal{B} + D\mathcal{H}$	N
Common Backbone	D	$1\mathcal{B} + D\mathcal{H}$	$1\mathcal{B} + D\mathcal{H}$	N
Single Model (Ours)	1	$1\mathcal{B} + 1\mathcal{H} + 1E_Q$ ($E_Q << \mathcal{H}$)	$1\mathcal{B} + 1\mathcal{H} + 1E_Q$ ($E_Q << \mathcal{H}$)	$0 - \infty$

yield any number of specific facial landmark predictions at inference time. In Table 1, we compare the efficiency of the single model paradigm of our method with the separate model and common backbone paradigms executed by prior work, when trained on D number of unique "N-point" datasets. Visibly, our method is the *most efficient* since it is agnostic to D on all the considered factors, and is the *most versatile* since it offers demand-specific landmark throughput.

Our contributions and their benefits are summarized below:

1. We propose the Face Part-Anchored Landmark Positions (FPALPs), an intuitive representation of face landmarks that are evenly distributed on well-defined face part curves. The FPALP format is universal and allows for compatibility with all existing and future datasets.
2. To the best of our knowledge, our work using FPALPs is the first to enable, *without auxiliary dataset information*, Unified FLD: the ability of a model to be trained end-to-end on the fusion of multiple "N-point" datasets. We demonstrate increased model generalization by training on a larger, more diverse combined dataset offering higher landmark heterogeneity through the unification of various "N-point" formats.
3. We propose a novel FPALP-based landmark queried regressor to enable Dynamic FLD, i.e., unlimited on-demand landmark prediction without network retraining.
4. We demonstrate through extensive experiments that our work not only unlocks the numerous benefits of Unified Dynamic FLD but also matches or outperforms existing state-of-the-art methods on several benchmark datasets.

2 Related Work

Targeting Fundamental Performance Improvements. Recent face landmark detection (FLD) methods can be categorized into direct coordinate regression methods [18, 40, 20] and heatmap-based regression methods [10, 46, 17]. While each approach has advantages and disadvantages, they target different challenges to achieve performance improvements. AnchorFace [41] uses a split-aggregate strategy using anchor templates, while MCUDN [36] combines uncertainty-aware regression with a multi-expert collaborative module to tackle landmark uncertainty in occluded and large face poses. ADNet [10] and STAR Loss [46] address the semantic ambiguity in landmark annotations by suppressing the associated disentangled loss component for landmarks with an anisotropic distribution. DTLD [18] and SLPT [40] adaptively leverage the underlying inter-landmark structural relationship to improve localization performance, especially on occluded landmarks. Meanwhile, PIPNet [13] performs simultaneous heatmap regression and offset predictions to speed up inference while achieving competitive localization accuracy. Recently, PossLoss [47] aimed at aligning landmark peaks and emphasizing hard examples to improve the precision and robustness of the predicted landmarks. *Orthogonal to these efforts*, our work aims to deliver the aforementioned unified and dynamic FLD properties to induce *robustness and versatility at the system level*.

Approaches to Ameliorate the FLD Pipeline. Prior works have also surfaced the issues of immiscibility of the various "N-point" annotation schemes across datasets [39, 42], and the infeasibility to infer landmarks beyond those N defined by the training dataset [42, 5]. LAB [39] represented facial structure using 13 boundary lines, theorized that facial landmarks across datasets can be interpolated within these lines, and performed landmark regression using the common backbone paradigm. LDDMM-Face [42] maps landmarks on mean face templates to semantic boundary flows and predicts final landmarks via flow-wise deformation layers, with limited cross-annotation adaptation handled by affine alignment between source and target mean faces. FreeEnricher [11] refines interpolated landmarks along a face-part curve using contextual patches to predict boundary-aligned offsets, but because this enrichment is decoupled from the base landmark detector, its effectiveness depends heavily on the accuracy of the initial landmark predictions. Recently, CLD [5] proposed a pipeline that ingests a facial image and arbitrary 3D query locations on a canonical face shape to output the corresponding and possibly continuous 2D landmark coordinates. Although CLD could be trained with multiple datasets, its success is highly dependent on a large collection of densely annotated face datasets having 3D canonical landmark mappings. In contrast, our single model paradigm trains end-to-end with only sparsely annotated 2D landmark datasets and performs dynamic and direct inference to any arbitrary landmark format *without manual transformations*. Furthermore, our Face Part-Anchored Landmark Position-based landmark queries are *easily interpretable* and allows for unconstrained interaction with text-based or agentic downstream applications.

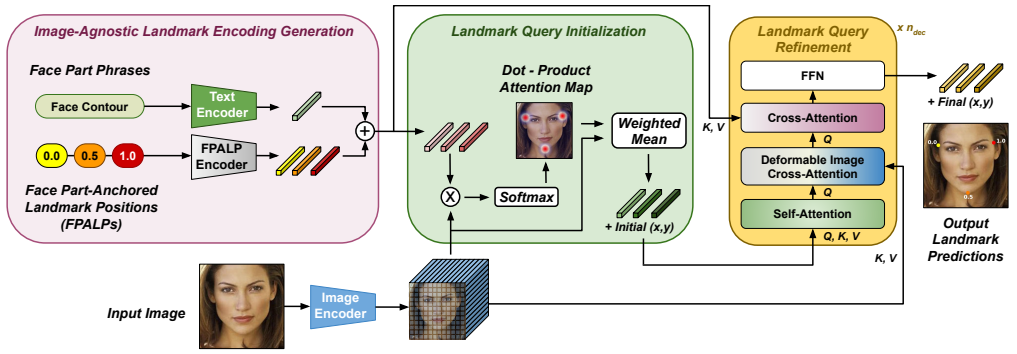


Figure 3: An overview of our proposed framework. First, we associate user-defined face parts to the required landmarks and calculate their Face Part-Anchored Landmark Positions (FPALPs). The FPALPs and the face parts' text are encoded and aggregated to yield the image-agnostic landmark encodings. The facial image's visual features are then conditioned on these encodings to output the initial landmark queries and coordinate predictions. Lastly, a cross-modality decoder block iteratively refines the landmark queries and coordinate predictions to output the final values.

Generalist Face Models. Another line of research aims to simultaneously perform facial tasks such as landmark detection, age/gender/head-pose estimation, and face parsing using multi-task learning. Early works like HyperFace [28] and AIO [29] utilized multi-scale features from various CNN layers and executed upto 7 face tasks at once using task-wise heads. Recently, FaceXFormer [24] and Faceptor [27] treated face tasks as tokens in transformer-based architectures [35] containing unified task and pixel decoders. These on-demand task-expandable generalist face models train on the fusion of diverse task datasets. However, for the FLD task, they still train separately on the "N-point" datasets and yield only a fixed N output. Our unified dynamic FLD method can be *readily integrated* into existing generalist face models to streamline their FLD division.

3 Methodology

Our proposed Unified Dynamic Face Landmark Detection framework is inspired by Grounding DINO [21] and is illustrated in Figure 3. First, we introduce Face Part-Anchored Landmark Positions (FPALPs), a supplementary representation of facial landmarks from the viewpoint of face part boundaries. Next, we describe how we construct image-agnostic landmark encodings using FPALPs and combine them with facial image features sourced from an image encoder to initialize the landmark queries and the primitive coordinate predictions. Lastly, we elucidate the process of iterative query refinement to yield the final landmark representations and coordinate predictions.

Face Part-Anchored Landmark Positions (FPALPs).

As prior work [39, 42] have noted and illustrated by us earlier in Figure 2a, facial landmarks specified by benchmark datasets that we consider, i.e., AFLW [48], WFLW [39], and 300W [31], are bound to face part boundaries in an evenly spaced manner. To leverage this observation, we conceptualize Face Part-Anchored Landmark Positions (FPALPs). Here, each landmark is associated with one or more containing face parts and is represented as a progression value between 0 and 1 denoting its fractional position within the containing face part curve. Since each "N-point" dataset can define its face template with different landmark layouts and different start and end positions for the various face parts, we first create a unified face template by taking the union of the face templates of all datasets. Formally, we denote the face template for dataset D_i, out of D considered datasets, having N_i number of landmarks, as T_{D_i}, and the unified face template as $T_{\mathrm{U}} = T_{D_1} \cup T_{D_2} \cup ... \cup T_{D_D}$. T_{U} consists of N_{U} number of landmarks clusters each of which indicates a landmark's proximity across the D datasets. While this may seem inexact, for the considered datasets, we observed a clean alignment between the face templates resulting in tight proximal landmark clusters having an average intra-cluster distance of 2.22 pixels averaged over all face parts.

We split T_{U} into P face-part templates, $T_{\mathrm{U}} = T_{P_1} \cup T_{P_2} \cup \cdots \cup T_{P_P}$, each consisting of member landmarks that represent user-defined face-part curves such as the left/right eye(brow), face contour, and inner/outer lip.

Face part curves can be open (e.g., nose bridge, face contour) or closed (e.g., eyes, lips). For closed curve face parts, we create a copy of the starting landmark in the curve sequence and signify it as the ending landmark for that sequence. In Figure 4, we exemplify the following FPALP formulation using the left eyebrow and right eye face part curves. For a landmark l positioned at $pos_{l,p}$ within a sequence of N_p landmarks that compose the face part p with template T_p, we define the FPALP of l as $FPALP_{l,p} = pos_{l,p}/(N_p - 1)$.

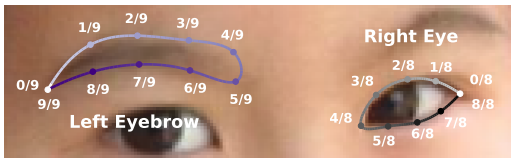


Figure 4: Illustration of the construction of Face Part-Anchored Landmark Positions (FPALPs) for the landmarks of the left eyebrow and the right eye closed-curve face parts.

Image-Agnostic Landmark Encodings. To achieve dynamic face landmark detection, we represent target landmarks as landmark queries. To this end, we construct initial image-agnostic representations which conceptually capture the landmarks to be queried. Firstly, we encode FPALPs using a simple MLP with ReLU activation. Next, we input the face part name into a lightweight pretrained text encoder to get its textual representation. Finally, we derive the image-agnostic landmark encodings as the summation of the encoded FPALPs and the face part textual representations. Formally, for a landmark l in the face part p, the image-agnostic landmark encodings $E^{l,p}_{\text{IA}}$ are derived as:

$$E^{l,p}_{\text{FPALP}} = \text{MLP}(FPALP_{l,p}), \quad E^{p}_{\text{text}} = \text{Enc}_{\text{text}}(p), \quad E^{l,p}_{\text{IA}} = E^{l,p}_{\text{FPALP}} + E^{p}_{\text{text}} \tag{1}$$

where $E^{l,p}_{\text{FPALP}}, E^{p}_{\text{text}}, E^{l,p}_{\text{IA}} \in \mathbb{R}^d$, and d is the encoding dimension. In lieu of Enc_{text}, we could use learnable embeddings to yield the face part representations. We hypothesize that pretrained text encoders are more superior since they may already encode the semantics of facial layouts. In Sec. 4.2, we compare both the options and corroborate that using pretrained text encoders is the better choice.

Landmark Query Initialization. Effective initial landmark queries should capture the required landmarks' proximity specified by their semantic definitions. To this end, we condition the facial image's visual features with the image-agnostic landmark encodings in the following manner. First, we utilize a pretrained image encoder to output the facial image features $E_{\text{I}} \in \mathbb{R}^{H_{\text{I}} \times W_{\text{I}} \times d}$, where $(H_{\text{I}}, W_{\text{I}})$ represents the spatial resolution of the image features. Let $G_{\text{I}} = \{(x^c_{i,j}, y^c_{i,j})\}^{H_{\text{I}}-1, W_{\text{I}}-1}_{i=0,j=0} \in \mathbb{R}^{H_{\text{I}} \times W_{\text{I}} \times 2}$ represent the grid of the image-space center coordinates corresponding to E_{I}. Next, we derive the attention map A of the visual features with respect to the required image-agnostic landmark encodings $E_{\text{IA}} \in \mathbb{R}^{L \times d}$, where L denotes the number of landmarks to be queried, as $A = \text{Softmax}(E_{\text{I}} \cdot E^T_{\text{IA}})$, where $A \in \mathbb{R}^{H_{\text{I}} \times W_{\text{I}} \times L}$ and the softmax is applied along the $H_{\text{I}} \times W_{\text{I}}$ dimension. Here, A reflects the activation of the visual regions that correspond to the required landmarks' image-agnostic landmark encodings. We obtain our initial landmark queries $Q_0 \in \mathbb{R}^{L \times d}$ and initial coordinate predictions $C_0 \in \mathbb{R}^{L \times 2}$ by taking the weighted mean of E_{I} and G_{I} using the attention map A, respectively. Formally, given grid center coordinates $(x^c_{i,j}, y^c_{i,j}) \in G_{\text{I}}$, for a required landmark $l \in [0, L)$ with a corresponding attention map $A^l \in \mathbb{R}^{H_{\text{I}} \times W_{\text{I}}}$, the initial landmark query LQ^l_0 and initial coordinate prediction C^l_0 are derived as:

$$LQ^l_0 = \sum_{i=0,j=0}^{H_{\text{I}}-1, W_{\text{I}}-1} A^l_{i,j} \cdot E_{\text{I}_{i,j}}, \qquad C^l_0 = \left(\sum_{i=0,j=0}^{H_{\text{I}}-1, W_{\text{I}}-1} A^l_{i,j} \cdot x^c_{i,j}, \sum_{i=0,j=0}^{H_{\text{I}}-1, W_{\text{I}}-1} A^l_{i,j} \cdot y^c_{i,j} \right) \tag{2}$$

We supervise A^l with the PossLoss [47] following the standard practice of transforming ground-truth landmarks into 2D Gaussian heatmaps.

Landmark Query Refinement. We employ a cross-modality transformer decoder, as depicted in the third block of Figure 3, to iteratively hone the landmark queries and the predicted coordinates. This block consists of n_{dec} decoder layers, the first of which consumes LQ_0 and C_0, while the later layers consume the output of the previous layers to implement iterative refinement. At layer dec_i, we first execute self-attention on the landmark queries $LQ_{\text{dec}_i - 1}$ to exploit the inter-landmark dependencies. Then, we deploy a deformable attention [49] layer that consumes the locations $C_{\text{dec}_i - 1}$ and performs targeted cross-modality attention between the image features and the queries from the previous step. To better align the queries with the semantic definitions of the target landmarks, we

apply a cross-attention layer between the previous-step queries and the image-agnostic landmark encodings E_{IA}. Finally, we deploy a feed-forward network to yield the decoder layer's query output LQ_{dec_i}, operate an MLP on it to derive the coordinate offsets with respect to C_{dec_i-1}, and calculate the coordinate predictions as C_{dec_i}. For brevity, we assume familiarity with standard transformer notation and describe the process below using simplified equations:

$$LQ^{SA}_{\text{dec}_i} = \text{SelfAttn}(LQ_{\text{dec}_i-1}, LQ_{\text{dec}_i-1}, LQ_{\text{dec}_i-1}) \tag{3}$$

$$LQ^{DICA}_{\text{dec}_i} = \text{DeformableAttn}(LQ^{SA}_{\text{dec}_i}, E_{\text{I}}, E_{\text{I}}, C_{\text{dec}_i-1}) \tag{4}$$

$$LQ^{CA}_{\text{dec}_i} = \text{CrossAttn}(LQ^{DICA}_{\text{dec}_i}, E_{\text{IA}}, E_{\text{IA}}) \tag{5}$$

$$LQ_{\text{dec}_i} = \text{FFN}(LQ^{CA}_{\text{dec}_i}) \tag{6}$$

$$C_{\text{dec}_i} = C_{\text{dec}_i-1} + \text{MLP}(LQ_{\text{dec}_i}) \tag{7}$$

where $\text{dec}_i \in [1, n_{\text{dec}}]$ and the first three inputs to the layers in Equation 3-5 respectively assume the roles of query, key, and value in the attention mechanism.

Given the ground truth coordinates of the required L landmarks $C_{\text{GT}} \in \mathbb{R}^{L\times 2}$, we supervise both our intermediate and final coordinate predictions C_{dec_i} where $\text{dec}_i \in [0, n_{\text{dec}}]$ using the Wing Loss [7] as $\mathcal{L} = \sum_{\text{dec}_i=0}^{n_{\text{dec}}} \text{WingLoss}(C_{\text{dec}_i}, C_{\text{GT}})$.

In Sec. 4.1, to estimate the maximum achievable performance on each dataset D_i, we inject LoRA [9] modules separately for each dataset, only into the final decoder layer ($\text{dec}_i = n_{\text{dec}}$), specifically its cross-attention and FFN sublayers. We refer to these dataset-specific LoRA modules as *Dataset Adapters* and use them to evaluate the corresponding dataset-specific fine-tuned network.

4 Experiments

Datasets. We train and evaluate our framework on three benchmark datasets: AFLW [48], 300W [31], and WFLW [39]. AFLW focuses on coarse annotations for in-the-wild images and comprises of 20000 training and 4386 test facial images, each annotated with 19 landmarks. 300W is collected from five facial datasets and contains 3148 training and 689 test facial images, each annotated with 68 landmarks. The test set is further divided into common (554 images) and challenging (135 images) subsets. WFLW is collected from WIDER Face [43] with an emphasis on challenging poses, expressions, and occlusions. It consists of 7500 training and 2500 test images, each annotated with 98 landmarks. For cross-dataset evaluation, we consider COFW [3], which contains 507 test images each annotated with 29 landmarks, COFW68 and WFLW68, the 68 landmark variants whose face template matches that of 300W. Collectively, these datasets provide images with diverse levels of expression, pose, and occlusion, making them effective to evaluate a model's generalization ability.

Implementation Details. Facial images from all datasets are cropped using the given bounding boxes and resized to either 224×224 (ViT-B) or 256×256 (ResNet) depending on the image encoder. Following prior works [13, 18, 27], bounding boxes are enlarged by 10% to include more contextual information. Data augmentation methods including random rotations ($\pm 15°$), scaling ($\pm 20\%$), horizontal flipping, and translation (± 10 pixels), are employed to improve model robustness by simulating real-world variability.

We employ the pretrained SentenceBERT [30] as the face part text encoder, FaRL [45] pretrained ViT-B [6] or ResNet [8] as the facial image encoder, 3 decoder layers (n_{dec}), each with 8 attention heads, and a model-wide feature dimension $d = 256$. During image cross-attention, 4 features per head are sampled from each level of the image feature maps for each query. We train the full model on an NVIDIA A100 GPU (40GB) for 32 epochs, with a batch size of 16, using the Adam optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-5}. The learning rate is lowered to 10^{-5} from the 25th epoch. The image and text encoders are trained at a tenth of the running learning rate. For the PossLoss [47], the weighting and temperature parameters are set to 2 and 0.1 respectively.

Evaluation Metrics. Following prior works [13, 18, 40], we evaluate the face landmark detection methods using the Normalized Mean Error (NME) percentage. NME measures the L2 distance between the predicted and true landmarks and is normalized by either the inter-ocular distance (NME_{io}), which is used for evaluation on 300W and WFLW, or the diagonal distance of the facial bounding box (NME_{diag}), which is used for evaluation on AFLW. For WFLW, we also report the Failure Rate (FR_{10}) based on a threshold of 10% error.

Table 2: Comparison of our Unified Dynamic Face Landmark Detection approach with SOTA methods on the WFLW, 300W, and AFLW-19 datasets. We include the generalist approaches that use additional datasets from other tasks for reference purpose only. *We enable fused dataset training and dynamic landmark prediction* while matching/outperforming SOTA methods on the full version of the datasets. **Bold** and underline indicates the best and second best results respectively.

Method	Method Type	Trained w/ Additional Datasets	Fused Dataset Training	Dynamic Landmark Prediction	WFLW Full		300W			AFLW-19 Full
					NME_{io} ↓	FR_{10} ↓	Common NME_{io} ↓	Challenge NME_{io} ↓	Full NME_{io} ↓	NME_{diag} ↓
FaceXFormer [24]	Generalist	✓	✗	✗	-	-	2.66	4.67	3.05	-
Faceptor [27]		✓	✗	✗	4.03	-	2.52	4.25	2.86	0.95
PIPNet [13]	Specialist	✗	✗	✗	4.31	-	2.78	4.89	3.19	1.42
ADNet [10]		✗	✗	✗	4.14	2.72	2.53	4.58	2.93	-
SLPT [40]		✗	✗	✗	4.14	2.76	2.75	4.90	3.17	-
DTLD+ [18]		✗	✗	✗	4.05	2.76	2.60	4.48	2.96	1.37
STAR Loss [46]		✗	✗	✗	**4.02**	2.32	2.52	4.32	2.87	-
PossLoss [47]		✗	✗	✗	4.07	**2.12**	2.51	4.21	2.84	-
MCUDN [36]		✗	✗	✗	4.15	-	**2.42**	4.33	2.79	1.47
Ours (ViT-B)	Specialist	✗	✓	✓	4.05	2.38	2.47	4.25	2.80	1.02
+ Dataset Adapters		✗	✓	✓	**4.02**	2.19	2.43	**4.19**	**2.76**	**1.01**

4.1 Comparison with SOTA Methods

Our main contribution is *the enablement of the unified and dynamic FLD features* rather than dataset-specific specialization. Accordingly, in the below subsection, we compare against prior methods both with and without Dataset Adapters. In all other experiments, Dataset Adapters *are not used*, since they are intended to assess the unified model itself. Since most prior methods are not open-sourced and often use different backbones, fair direct comparison under a common setting is challenging. Hence, we cite their reported results and explicate our model configuration for transparency.

Individual Dataset Evaluation. We compare our Unified Dynamic Face Landmark Detection (FLD) framework with SOTA methods in Table 2. Without Dataset Adapters, our model enables joint training on multiple datasets and supports dynamic landmark prediction, while achieving performance on par with prior state of the art. This points to a *well-defined alignment* among the face templates of the considered datasets and demonstrates that Face Part-Anchored Landmark Positions (FPALPs) constitute an effective representation for FLD. With Dataset Adapters, performance improves further and consistently surpasses prior methods across datasets. This indicates that the model learns a robust unified representation of landmark semantics, from which *dataset-specific offsets can be effectively recovered* through lightweight adaptation.

Cross-Dataset Evaluation. To verify the generalization ability of our approach, we conduct a cross-dataset evaluation on the COFW68 and WFLW68 datasets using our model trained only on the 300W dataset, and present the results in Table 3. Our method with the ResNet backbones fare approximately on par with SOTA on the 300W and COFW68 datasets. Using the ViT-B backbone, we demonstrate robustness by significantly improving performance on the challenging WFLW68 dataset, which includes facial images with extreme poses, expressions, occlusions, and makeup.

Table 3: Cross-dataset evaluation comparison. Models are supervised only on 300W.

Method	300W	COFW68	WFLW68
	NME_{io} ↓		
LAB[39]	3.49	4.62	-
AVSw/SAN[26]	3.86	4.43	-
DAG[20]	3.04	4.22	-
PIPNet[13]	3.36	4.55	8.09
DTLD[18]	3.07	4.42	7.23
Ours (ResNet18)	3.10	4.67	7.11
Ours (ResNet101)	3.05	4.59	6.86
Ours (ViT-B)	**3.01**	**4.40**	**6.08**

4.2 Ablation Studies

Our model is trained on a fusion of AFLW19, 300W, and WFLW, and we conduct a dataset ablation study to assess the contribution of each dataset to overall performance (Table 4). The study evaluates the impact of training on salient dataset combinations and tests on individual datasets. Notably, when the evaluation dataset uses a different "N-point" template than those seen during training, the setting is effectively *near zero-shot*, as many FPALPs in the target dataset are unseen. Our approach is the first to conduct such cross-template evaluations without resorting to manual interpolation techniques. The results indicate that training on all datasets combined yields the best performance across most datasets, except for WFLW68, where the best performance is achieved by training solely

Table 4: Ablation study on training datasets as a cross-dataset evaluation. * indicates exclusion of undefined landmarks not defined in the template.

Training Datasets	AFLW-19 NME_{diag} ↓	300W	WFLW	COFW	WFLW68	COFW68
		NME_{io} ↓				
300W	2.18*	3.01	6.32*	3.81*	6.08	4.40
WFLW	2.21*	4.03	4.09	3.71	**3.89**	4.61
300W + WFLW	2.20*	2.89	4.11	3.64	4.41	4.36
300W + WFLW + AFLW-19	**1.02**	**2.80**	**4.05**	**3.52**	4.38	**4.27**

on WFLW. We attribute this to the reduction of non-critical landmarks in the transition from the 98-point to the 68-point template and the dilution of challenging samples when additional datasets are introduced. It is essential to acknowledge that the effect of incorporating new training datasets can vary based on alignment between the distribution and label quality of the training and evaluation datasets. The observed gains from training on datasets with diverse face templates suggest that exposure to varied FPALPs enhances the model's ability to effectively represent face part curves and generalize across varied facial structures and ambient conditions.

Choices for Face Part Representation. In this study, we investigate the impact of choosing how face parts are represented to yield E^{p}_{text} in Equation 1. We present two options: (1) training learnable embeddings, or (2) leveraging the output of pretrained text encoders. The former might seem as the default option given the limited amount of face parts that can be encoded. We contend that, although simpler, training with *learnable embeddings may not capture the nontrivial semantics of facial structure*, such as the relative positions of face parts, the inter-face part relationships during facial expressions (e.g., the squinting of the eyes and broadening of the lips during a laugh), and interactions with makeup and accessories. We postulate that text encoders that are trained on diverse corpora encode these intricacies. In Figure 5, we compare the training curve plots of the model when using learnable embeddings versus SentenceBERT [30], a lightweight pretrained text encoder, and in Table 5, we compare their performance at convergence. The usage of SentenceBERT to represent face parts results in a faster convergence and a more performant model, thereby corroborating our earlier thesis and proving to be the superior choice over learnable embeddings.

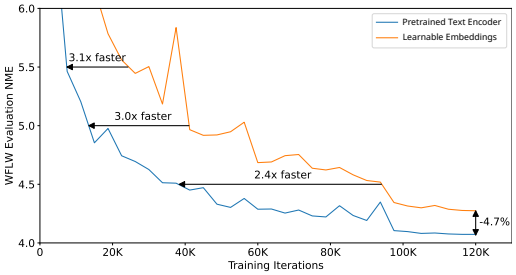


Figure 5: A comparison of training curve plots when using learnable embeddings vs. SentenceBERT [30] to represent face parts.

Table 5: Performance comparison using learnable embeddings vs. SentenceBERT [30].

Query Type	AFLW-19 $\text{NME}_{\text{diag}} \downarrow$	300W	WFLW	COFW	COFW68
		$\text{NME}_{\text{io}} \downarrow$			
Learnable	1.07	2.99	4.19	3.73	4.35
Language	**1.02**	**2.80**	**4.05**	**3.52**	**4.27**

Impact of Image Encoder. We analyze the impact of different image encoders, including ResNet18, ResNet101, and ViT-B, on the overall performance of our model as well as its generalization ability. In Table 6a, we display the results of our model with different backbones when trained on the fusion of the considered datasets. Although ViT-B proves to be superior on most of the evaluation datasets, it is noteworthy that both ResNets perform competitively, at a fraction of the size of ViT-B. This suggests that the availability of diverse "N-point" training datasets is of higher importance than the capacity of the image encoder to achieve an overall high-performing model. In Table 6b, we present the results of a cross-dataset evaluation conducted by training our model only on 300W, with different image encoders, and report its performance on 300W, COFW68, WFLW, WFLW68, and WFLW_E, whose face template contains only the 28 points that are absent/undefined in 300W. We observe that as the image encoder's capacity increases, the performance improves drastically on most datasets, especially on the WFLW_E variant where our model executes *zero-shot* evaluation since the landmarks (and their corresponding FPALPs) are *unseen* during training. As we constrained our model training to only 300W, the result of this experiment suggests that the generalization ability of our model is dependent on the capacity of the image encoder.

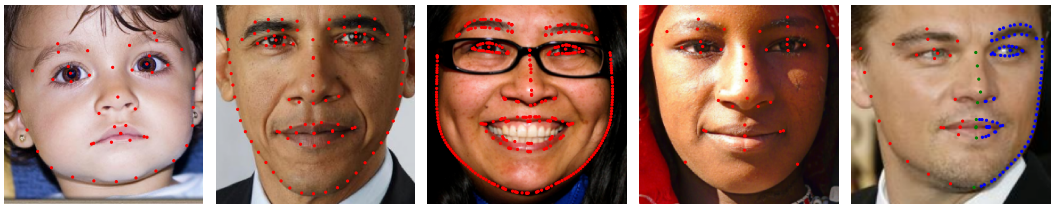
Table 6: (a) Performance comparison of our model trained on the fusion of all considered datasets when using various image encoders. (b) Cross-dataset evaluation with our model trained only on 300W when using different image encoders. WFLW_E refers to the WFLW dataset containing *only the 28 points* that are absent/undefined in 300W.

Image Encoder	AFLW-19 $\text{NME}_{\text{diag}} \downarrow$	300W	WFLW	COFW	COFW68
		$\text{NME}_{\text{io}} \downarrow$			
ResNet18	1.05	2.97	4.36	3.73	4.50
ResNet101	1.03	2.84	4.19	3.65	4.42
ViT-B	**1.02**	**2.80**	**4.05**	**3.52**	**4.27**

(a)

Method	300W	COFW68	WFLW68	WFLW	WFLW_E
	$\text{NME}_{\text{io}} \downarrow$				
ResNet18	3.10	4.67	7.11	7.48	7.63
ResNet101	3.05	4.59	6.86	7.09	7.39
ViT-B	**3.01**	**4.40**	**6.08**	**6.32**	**6.52**

(b)



(a) (b) (c) (d) (e)

Figure 6: An illustration of the dynamic landmark prediction capability of our system. Images are selected from the WFLW (98-point) test set. We anchor the landmarks to the following face parts: left and right eyes, eyebrows, and pupils, inner and outer lips, face contour, nose bridge and boundary. For (e), we split the face contour, lips, and nose boundary into left, center, and right sub-parts. Landmark predictions per face part are depicted using (a)-(c) a granularity multiplier of 0.5, 1, and 4 respectively, (d) 4 landmarks per face part, and (e) a granularity multiplier of 0.5, 1, and 2 for left, center, and right sub-parts whose landmarks are color coded as red, green, and blue respectively.

4.3 Dynamic Landmark Prediction

In this section, we qualitatively assess the outputs of our Unified Dynamic Face Landmark Detection system. Using the model trained on the fusion of the considered datasets, we present a variety of dynamic landmark prediction configurations on images from the WFLW test set in Figure 6, and on challenging images with large pose variations and occlusions in Appendix Figure 8-9. In contrast to prior works, which can only output a fixed landmark layout as in Figure 6b, our model can predict landmarks pertaining to user-selected face parts and, furthermore, at various granularities within (Figure 6a-6d) and across (Figure 6e) face parts, hence demonstrating its versatility and applicability to an assortment of downstream applications. We observe that landmark predictions for face parts with higher FPALP diversity, such as the face contour and the eyes, are more accurate than those with lower FPALP diversity, such as the nose boundary. A larger diversity of face templates within the combined training dataset increases the exposure to different FPALPs and empowers the model to accurately predict landmarks at higher granularities. Additionally, incorporating loss components that enforce appropriate distribution of landmark predictions across the input FPALPs could further enhance prediction quality – a direction we leave for future work.

5 Discussion

Limitations. We acknowledge the likelihood of imprecise alignment of the individual datasets' face templates during construction of the unified face template, which may hinder scalability of the FPALP formulation. We note from the resiliency of our model, trained on the alignment of three different (14, 68, and 98-point) face templates, that only approximate alignment is necessary for effective face part curve learning. In cases of large misalignment, new face parts can be defined to contain the introduced landmarks. Meaningful inter-face part relationships with the misaligned face parts can still be forged via the landmark query refinement process in our model.

Future Work. Future work could extend FPALPs to 2D representations that capture face-part surfaces and use query features to track facial artifacts such as acne, moles, and wrinkle lines. Integrating FPALPs with vision-language and generalist face models may further enable text- or visual-prompt-based face-part creation, automated landmark-to-FPALP registration, and unified template generation, yielding a more versatile face landmark detection component within a robust face analysis system.

6 Conclusion

In this paper, we present our Unified Dynamic Face Landmark Detection method, wherein landmarks are treated as progression points on user-defined face parts, allowing for end-to-end model training on the fusion of diverse "N-point" datasets and execution of unlimited on-demand landmark predictions. With performance competitive with, if not surpassing, SOTA methods, our simple yet adaptable framework is positioned to meet the requirements of various downstream applications that depend on a wide range of precise face landmarks.

References

[1] Arqam M. Al-Nuimi and Ghassan J. Mohammed. Face direction estimation based on mediapipe landmarks. In *ICCITM*, 2021.

[2] F. Bougourzi, F. Dornaika, and A. Taleb-Ahmed. Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. *Knowledge-Based Systems*, 2022.

[3] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[4] Mingcheng Cai, Shuo Zhang, Guoqiang Xiao, and Shoucheng Fan. 3d face reconstruction and dense alignment with a new generated dataset. *Displays*, 2021.

[5] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, and Derek Bradley. Continuous landmark detection with 3d queries. In *CVPR*, 2023.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, 2018.

[8] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[10] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *ICCV*, 2021.

[11] Yangyu Huang, Xi Chen, Jongyoo Kim, Hao Yang, Chong Li, Jiaolong Yang, and Dong Chen. Freeenricher: Enriching face landmarks without additional cost. *AAAI*, 2023.

[12] Tharun J. Iyer, Rahul K., Ruban Nersisson, Zhemin Zhuang, Alex Noel Joseph Raj, and Imthiaz Refayee. Machine learning-based facial beauty prediction and analysis of frontal facial images using facial landmarks and traditional image descriptors. *Computational Intelligence and Neuroscience*, 2021.

[13] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *IJCV*, 2021.

[14] Yi Jin, Xingyan Guo, Yidong Li, Junliang Xing, and Hui Tian. Towards stabilizing facial landmark detection and tracking via hierarchical filtering: A new method. *Journal of the Franklin Institute*, 2020. ISSN 0016-0032.

[15] Aniwat Juhong and C. Pintavirooj. Face recognition based on facial landmark detection. In *BMEiCON*, 2017.

[16] Robin Kips, Ruowei Jiang, Sileye Ba, Edmund Phung, Parham Aarabi, Pietro Gori, Matthieu Perrot, and Isabelle Bloch. Deep graphics encoder for real-time video makeup synthesis from example. In *CVPRW*, 2021.

[17] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, 2020.

[18] Hui Li, Zidong Guo, Seon-Min Rhee, Seungju Han, and Jae-Joon Han. Towards accurate facial landmark detection via cascaded transformers. In *CVPR*, June 2022.

[19] TianXing Li, Zhi Yu, Edmund Phung, Brendan Duke, Irina Kezele, and Parham Aarabi. Lightweight real-time makeup try-on in mobile browsers with tiny cnn models for facial tracking. In *CVPRW*, 2019.

[20] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, et al. Structured landmark detection via topology-adapting deep graph learning. In *ECCV*, 2020.

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2023.

[22] Davide Marelli, Simone Bianco, and Gianluigi Ciocca. Designing an AI-based virtual try-on web application. *Sensors*, 2022.

[23] M. I. N. P. Munasinghe. Facial expression recognition using facial landmarks and random forest classifier. In *ICIS*, 2018.

[24] Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. *arXiv preprint arXiv:2403.12960*, 2024.

[25] Quang Tran Ngoc, Seunghyun Lee, and Byung Cheol Song. Facial landmark-based emotion recognition via directed graph neural network. *Electronics*, 2020.

[26] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *ICCV*, 2019.

[27] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. In *ECCV*, 2024.

[28] Rajeev Ranjan, Vishal Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *TPAMI*, 2016.

[29] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *FG*, 2017.

[30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

[31] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, June 2013.

[32] Divija Sanapala and J. Angel Arul Jothi. Virtual makeup try-on system using cognitive learning. In *Cognitive Analytics and Reinforcement Learning*, 2024. ISBN 9781394214068.

[33] Adil Sarsenov and Konstantin Latuta. Face recognition based on facial landmarks. In *AICT*, 2017.

[34] Yasmina Souley Dosso, Daniel Kyrollos, Kim Greenwood, Joann Harrold, and James Green. Nicuface: Robust neonatal face detection in complex nicu scenes. *IEEE Access*, 2022.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[36] Jun Wan, Hui Xi, Yuanzhi Yao, Hang Sun, Zhihui Lai, and Jie Zhou. Interpretable facial landmark detection by multi-expert collaborative uncertainty-aware deep networks. *Neural Networks*, 2026.

[37] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, Tom Cashman, and Julien Valentin. 3d face reconstruction with dense landmarks. In *ECCV*, 2022.

[38] Bing-Fei Wu, Bo-Rui Chen, and Chun-Fei Hsu. Design of a facial landmark detection system using a dynamic optical flow approach. *IEEE Access*, 2021.

[39] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018.

[40] Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang, and Min Xu. Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In *CVPR*, 2022.

[41] Zixuan Xu, Banghuai Li, Miao Geng, Ye Yuan, and Gang Yu. Anchorface: An anchor-based facial landmark detector across large poses. In *AAAI*, 2020.

[42] Huilin Yang, Junyan Lyu, Pujin Cheng, Roger Tam, and Xiaoying Tang. Lddmm-face: Large deformation diffeomorphic metric learning for cross-annotation face alignment. *Pattern Recognition*, 2024. ISSN 0031-3203.

[43] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016.

[44] Ruiqi Zhao, Tianyi Wu, and Guodong Guo. Sparse to dense motion transfer for face image animation. In *ICCVW*, 2021.

[45] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, June 2022.

[46] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. STAR Loss: Reducing semantic ambiguity in facial landmark detection. In *CVPR*, 2023.

[47] Qikui Zhu. Possloss: A reliable and sensitive facial landmark detection loss function. In *ICCV*, 2025.

[48] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.

[49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*, 2021.

A Technical Appendices and Supplementary Material

A.1 Relevance and Strength of Contribution

Our work addresses critical limitations in current 2D face landmark detection (FLD) methods and provides an efficient, semantically flexible alternative to both dense and traditional sparse approaches:

- **Unified Training Across Datasets.** We introduce the FPALP representation, which enables a single model to be trained across heterogeneous landmark templates without requiring 3D priors or costly alignment procedures. This overcomes the fragmentation seen in prior work, where separate models are typically needed for different datasets.
- **Dynamic, Semantic Landmark Prediction.** Unlike fixed-protocol models or dense outputs, our method supports flexible, part-based landmark queries. This design offers interpretability and adaptability for downstream tasks that demand only specific landmarks or face regions.
- **Purely 2D Supervision.** Our method operates entirely within the 2D domain, without relying on 3D annotations or model-based priors. This makes it more scalable and accessible in real-world applications where 3D data is limited or unavailable.
- **Improved Generalization and Regularization.** Training on diverse datasets with different landmark configurations serves as a natural regularizer, promoting robustness and reducing overfitting. The FPALP structure aligns these heterogeneous protocols into a unified representation that supports generalization to unseen templates.
- **Compatibility with Sparse-to-Dense Learning.** Our model supports zero-shot or near-zero-shot generalization across protocols. It can be trained on sparse landmarks and still perform well on denser configurations, laying the groundwork for bridging sparse and dense paradigms in a single framework.
- **Suitability for Low-Resource Deployment.** Dynamic 2D FLD is particularly advantageous for edge and mobile devices, where lightweight, semantically interpretable, and adaptable models are essential. Our method meets these requirements, while 3D-based approaches—due to their reliance on dense meshes, heavy computation, and 3D priors—are ill-suited for such environments. By avoiding these constraints, our framework provides a practical and efficient solution for real-world deployment.

A.2 Limitations Continued

From our ablation studies, we infer that the generalization ability of our model, which we define as its capacity to accurately predict landmarks at unseen Face Part-Anchored Landmark Positions (FPALPs) – is influenced by both the diversity of training dataset face templates and the range of facial and ambient conditions. Training on a broader variety of datasets with distinct landmark layouts, rather than simply increasing the number of datasets with similar layouts, is likely to yield a more generalizable model. However, as discussed in the dataset ablation section, incorporating additional training datasets may enhance overall generalization but could also diminish performance on specific evaluation datasets if the training and evaluation datasets differ significantly in facial and ambient condition distributions or label quality.

We also note that FPALPs are constructed using native dataset annotations. These annotations along face part boundaries often represent semantic progression points in 2D space. However, landmarks generated through interpolation techniques may not align with those predicted via evenly spaced FPALPs. For instance, consider a front-facing view of a person whose jawline narrows near the chin. Landmarks sampled along the true jaw contour may increase in density as they approach the chin. While these landmarks may appear equidistant in the profile view, they may not be evenly spaced in the frontal view. Consequently, we are unable to quantitatively evaluate our method on benchmarks that employ higher-density landmarks derived through interpolation, as it becomes challenging to objectively identify the cases we described above.

Lastly, we acknowledge that the choice of the training dataset and the text encoder could introduce biases or limitations when face part phrases are described using low-resource languages. In order to mitigate such biases and limitations, the definition of face part phrases should be standardized for a consistent interpretation across the different languages, and the text encoder would either need to be trained or fine-tuned on the target languages. While this is a crucial consideration for real-world

deployment in a global context, our current work limits its scope to the English language context, deferring inter-linguistic adaptations and broader cross-cultural considerations to future research.

A.3 Exclusion of Datasets

Our framework is trained and evaluated on AFLW19 [48], WFLW [39], and 300W [31], with additional evaluation conducted on COFW [3] and its variants. However, we exclude COFW from the training set due to observed inconsistencies in annotation quality. Preliminary experiments indicated that including COFW not only degraded the overall performance across all datasets but also adversely impacted the quality of denser landmark predictions. We also do not evaluate on the Enriched 300W test set proposed in [11] as its annotations are derived through interpolation-based methods, as discussed in the subsection A.2 above.

A.4 Choice of Text Encoder

As discussed in the main paper, we employ SentenceBERT [30] as the text encoder to generate face part representations. In Sec. 4.2, we detailed the rationale for selecting a language model output rather than a learnable embedding. Another plausible option was to use the FaRL [45] text encoder, given that we already utilize its image encoder. Although FaRL was trained on LAION-FACE [45], a dataset comprising facial image-text pairs, the textual descriptions predominantly consist of general attributes such as “smiling girl with party wig" or “the beautiful bride with the sunlight shining on her," rather than the specific face part intricacies discussed in Sec. 4.2. In contrast, SentenceBERT, having been pretrained on diverse and extensive textual corpora, demonstrated a superior ability to effectively encode these more detailed and nuanced characteristics of individual face parts.

While our study confirmed that using the FaRL text encoder yielded superior performance compared to generic learnable embeddings, our experiments ultimately revealed that SentenceBERT outperformed FaRL’s text encoder for our specific task. This indicated that, for generating image-agnostic landmark encodings using face part phrases, a superior representation of the specific semantics of face parts is achieved by using a strong pretrained text encoder. Therefore, despite the potential benefits of image-text alignment, the pretrained, lightweight SentenceBERT proved to be the more effective choice for encoding face part phrases in our framework.

A.5 Implementation Details for Fine-Tuning with Dataset Adapters

In Sec. 4.1, we leverage the Dataset Adapters which was described in Sec. 3 to report the maximum performance we can achieve with our approach for each dataset. We first train our model with the unified dataset and freeze the network. Then, for each dataset, we attach and train *only* the Dataset Adapters, consisting of LoRA [9] modules of rank 4 to the cross-attention and FFN layers of *only the last* decoder block, for 5 epochs at a learning rate of 10^{-5}.

A.6 Quantitative Evaluation of Dynamically Queried Landmarks

We further evaluate whether the proposed framework can accurately predict landmark positions that are not directly observed during training. This is particularly important for validating the dynamic landmark prediction capability beyond the qualitative examples presented in the main paper.

Existing cross-template evaluation. The main paper already provides two evaluations involving unseen landmark queries. First, Table 4 evaluates models on landmark templates that differ from those used for training. For example, models trained only on 300W or WFLW are evaluated on AFLW-19, WFLW, COFW, WFLW68, and COFW68. Whenever the evaluation template differs from the training template, a subset of the requested FPALPs is unseen during training, providing a near-zero-shot evaluation of cross-template landmark prediction.

Second, Table 6b isolates unseen landmark positions more directly. A model trained only on the 68-point 300W template is evaluated on WFLW_E, which contains the 28 WFLW landmarks that are absent from 300W. Despite never observing these FPALPs during training, our ViT-B model achieves an NME of 6.52, with performance improving systematically as backbone capacity increases.

Table 7: Held-out landmark prediction compared with cubic-spline interpolation. Full supervision is trained using the complete landmark template, while reduced supervision observes only the retained landmarks. Cubic splines are fitted to the retained landmark predictions of the reduced-supervision model. Lower NME is better.

Dataset and split	Full supervision	Reduced supervision	Cubic spline
300W: 37 retained, 31 held out	**3.56**	4.08	5.04
WFLW: 73 retained, 23 held out	**4.51**	4.92	5.88
WFLW: 50 retained, 46 held out	**5.32**	5.76	6.83

Controlled held-out landmark prediction. To more directly quantify the ability to predict unseen landmark positions, and to compare against geometric interpolation, we construct a controlled held-out landmark experiment using the native annotations of 300W and WFLW. Since these datasets do not provide ground-truth annotations for arbitrarily increased landmark granularity, we instead withhold a subset of their annotated landmarks during training and evaluate prediction accuracy exclusively on these unseen positions.

For each face-part contour, we retain landmarks by alternating landmark indices within the face-part definitions given in Table 12 and Table 13 of subsection A.15. For face parts containing an odd number of landmarks, the additional landmark is retained. This produces the following splits:

- **300W-50%:** 37 of 68 landmarks are retained for training, while the remaining 31 landmarks are held out for evaluation.
- **WFLW-75%:** after excluding the two eye-center landmarks, which do not lie on interpolatable contours, 73 of the remaining 96 landmarks are retained and 23 are held out.
- **WFLW-50%:** 50 landmarks are retained and 46 are held out.

We evaluate three settings:

1. **Full supervision:** the model is trained using the complete 68- or 98-point landmark template and evaluated only on the designated held-out subset. This provides a fully supervised reference.
2. **Reduced supervision:** the model is trained using only the retained landmarks and is directly queried at the FPALPs corresponding to the held-out landmarks.
3. **Cubic-spline interpolation:** cubic splines are fitted independently for each face part using the retained landmark predictions from the reduced-supervision model and are sampled at the FPALPs corresponding to the held-out landmarks.

We use cubic rather than linear interpolation because facial landmarks predominantly describe smooth, nonlinear contours. Following sparsification, some face parts contain as few as three retained landmarks, for which linear interpolation would produce piecewise-linear boundaries that cannot capture the natural curvature of structures such as the eyes, lips, eyebrows, and jawline. For closed contours, we use periodic cubic splines to preserve smoothness and continuity across the closure point. This therefore provides a strong contour-aware interpolation baseline rather than naive interpolation between neighboring landmarks.

Table 7 reports NME computed exclusively over the held-out native landmarks. Under identical reduced supervision, directly querying the proposed model at unseen FPALPs improves over cubic-spline interpolation by **19.0%** on the 37/31 300W split, **16.3%** on the 73/23 WFLW split, and **15.7%** on the 50/46 WFLW split. The reduced-supervision model also remains relatively close to the fully supervised reference despite never observing the evaluated FPALPs during training.

These results indicate that the proposed formulation learns more than a geometric interpolation rule between neighboring landmarks. The spline baseline is completely determined by the retained landmark predictions and consequently propagates localization errors from these sparse anchor points to the interpolated locations. In contrast, each queried landmark in our framework is predicted directly from the input image using its semantic face-part representation, FPALP encoding, and learned inter-landmark relationships.

This distinction is particularly relevant for challenging samples involving profile faces, partial occlusions, extreme poses, or strong facial expressions. For simple frontal faces, smooth contour interpolation may often be sufficient. Under more difficult conditions, however, interpolation cannot incorporate additional image evidence at the queried position and inherits errors from the neighboring retained landmarks. Our image-conditioned formulation can instead exploit both local visual evidence and global facial structure when localizing each requested landmark.

A.7 Ablation Study on the Number of Decoder Blocks

In Table 8, we report an ablation study on the number of decoder blocks (n_{dec}) used in our final model. We vary n_{dec} from 1 to 5 and observe rapid performance gains when increasing the number of blocks from 1 to 3, followed by degradation starting at 4 blocks, suggesting overfitting. Based on this trend, we set $n_{\text{dec}} = 3$ in our final network configuration.

A.8 Detailed Comparison with Continuous Landmark Detection [5]

As outlined in Sec. 2, Continuous Landmark Detection (CLD) [5] is a recent framework that takes as input a facial image and arbitrary 3D query locations on a canonical 3D face surface to output corresponding 2D landmark predictions. While CLD can be trained using existing 2D face landmark datasets, it requires a layout mapping to the 3D canonical surface, imposing a dependency on such mappings. In contrast, our Unified Dynamic FLD framework eliminates this dependency by deriving FPALPs directly from the native coordinate system of the dataset, enabling training on all native 2D FLD datasets without additional mappings.

Furthermore, CLD leverages 3D coordinates on the canonical face mesh as input queries, facilitating continuous landmark detection, and is a very valuable contribution, especially in applications where accurate dense coordinates are required to retrieve and characterize a facial surface. In contrast, our framework is designed to provide a more interpretable and semantically driven interface for querying FLD systems. Specifically, we construct queries based on text-defined face parts and semantic progression points along face contours. We envisage future FLD systems being queried using descriptive instructions, such as "Predict 10 coordinates from the left chin boundary to the end of the jawline," and position our framework to address such needs.

From an architectural perspective, CLD's query encoder processes 3D query locations on a canonical face mesh, whereas our framework encodes face part text and FPALPs. Additionally, while CLD's landmark predictor employs transformer layers to fuse the image encoder output and the 3D query encoding, our framework instantiates the initial landmark queries and coordinate predictions by encoding the face part text and corresponding FPALPs and subsequently conditioning the image features. We then refine the queries and coordinate predictions using self-attention and cross-attention layers to produce the final landmark coordinates.

A.9 Detailed FPALP Formulation

In this section, we revisit and elaborate on the formulation of the Face Part-Anchored Landmark Positions (FPALPs). Referring to the FPALP formulation in the main paper, for a landmark l positioned at $pos_{l,p}$ in a sequence of N_p landmarks that composes the face part p with template T_p, we denote the FPALP of l as $FPALP_{l,p} = pos_{l,p}/(N_p - 1)$. As observed, the landmark layout pertaining to a face part defined by a dataset usually comprises of landmarks that are evenly distributed on the face part boundary. The unification of individual landmark templates of the various

Table 8: Ablation study on the number of decoder blocks. Results are reported on the full versions of WFLW, 300W, and AFLW-19.

Number of Decoder Blocks (n_{dec})	WFLW		300W	AFLW-19
	$\text{NME}_{\text{io}} \downarrow$	$\text{FR}_{10} \downarrow$	$\text{NME}_{\text{io}} \downarrow$	$\text{NME}_{\text{diag}} \downarrow$
1	4.41	2.69	3.12	1.24
2	4.18	2.44	2.99	1.08
3	**4.05**	**2.38**	**2.80**	**1.02**
4	4.07	**2.38**	2.81	1.04
5	4.06	2.41	2.82	1.07

datasets into T_{U} may render the collection of landmarks to be unevenly distributed along the face part boundary. To determine $pos_{l,p}$ of a landmark l which originally belonged to the dataset D_i with landmark layout T_{D_i}, we first derive the position of the face part's starting landmark in T_{D_i} relative to the starting landmark of the face part in T_{U}, and then add to it the index of l relative to the other landmarks of the face part in T_{D_i}. We express the above formulation for $pos_{l,p}$ as:

$$pos_{l,p} = \text{RelativePosition}(l_{start,p}^{T_{D_i}}, l_{start,p}^{T_{\mathrm{U}}}) + index_{l,p}^{T_{D_i}} \tag{8}$$

The FPALP formulation normalizes progression along each face part from 0 to 1 in a dataset-agnostic manner. Crucially, the start and end points of a face-part curve in FPALPs are not fixed by any dataset; they are defined once by the practitioner when specifying the face-part phrase and its member landmarks. FPALPs then assign each (face part, landmark) pair a normalized position in [0,1] along that user-defined ordering. This means that the same physical landmark can legitimately receive different FPALP values under different, possibly overlapping, face-part phrases (e.g., "nose" = "nose bridge + left + right boundary of nose" vs "right nasal region" = "right boundary of nose + right nasolabial fold"), and differences in how individual datasets choose their "first" or "last" landmark on a contour do not constrain the unified representation.

A.10 Handling Undefined or Occluded Face Parts

Our current framework assumes that the queried face parts are explicitly defined in the training data. We acknowledge that parts that are heavily occluded or undefined poses a challenge and the impact would be dependent on the extent of visible visual context. To address such cases, the framework could be extended in future work to dynamically infer or adapt face part boundaries:

1. We can utilize the text encoder to parse face part descriptions into latent embeddings that can be aligned with image features. Soft spatial attention maps based on the introduced face parts can be used to approximate the boundaries of unseen face parts, even under occlusion. Such an extension would enable the model to infer FPALP-like progression values for novel regions by projecting the learned attention map onto surrounding anchor contours. Additionally, a dynamic part discovery module could be trained using contrastive losses to bind new textual descriptions to consistent visual patterns across samples. This could potentially enable open-vocabulary part generalization in FLD, which could be an exciting avenue for future work.
2. We can also leverage visibility annotations per landmark, such as those provided in the MERL-RAV [17] dataset, to supervise the model in learning to selectively ignore occluded regions during training. This allows the framework to learn robust part representations even when portions of the face are not visible. Additionally, these visibility flags can be used to guide a gating mechanism or soft-attention masking module that modulates the contribution of occluded regions in the query or image features during inference, improving landmark prediction reliability under occlusion.

A.11 Detailed 2D FPALP Proposal for Future Work

As discussed in the future work section, the proposed FPALPs can be extended to 2D space to facilitate further advancements. Currently, FPALPs are defined in 1D space, representing semantic progression points along a face part curve. By treating face part curves as boundaries, 2D FPALPs can be defined along these boundaries, capturing semantic progression both horizontally and vertically, with either the x or y component as zero. Extending this further, regions within the face part boundary can be described using 2D FPALPs where both x and y components are non-zero. With only the face part boundary as input, weak supervision could be employed to predict 2D FPALPs for arbitrary points within the face part region. Thus, transitioning from 1D to 2D FPALPs shifts the representation from linearly traversing face part curves to encompassing face part surfaces.

While 1D FPALPs correspond to progression along a face part boundary, 2D FPALPs require a surface parameterization that maps internal face part regions to a normalized coordinate space. Constructing such mappings without dense annotations firstly requires us to define the boundary coordinates of each defined face part in both spatial dimensions and further necessitates the use of weak supervision to learn the face part surface. For example, given only the boundary of a region (e.g., the cheek or

forehead), one could generate pseudo-ground-truth 2D FPALP labels using mesh-based interpolation to learn consistent internal representations across identities.

Incorporating 2D FPALPs would allow the model to reason over continuous face surfaces rather than just boundary curves, enabling richer spatial representations. This would benefit tasks such as facial expression analysis, where subtle shape changes within a region (e.g., the bulging of cheeks or wrinkling of the forehead) may not be captured effectively through sparse boundary points. By modeling internal face part regions with 2D FPALPs, the framework could localize and track deformations more precisely, potentially improving performance on downstream tasks requiring dense spatial awareness.

A.12 Detailed Training Procedure

Dataset Sampling. As our model is trained on a fusion of multiple datasets, we apply dataset-level oversampling to ensure a balanced training distribution. Each training epoch includes approximately the same number of samples from each dataset, ensuring equal exposure to each N-point facial landmark template.

Batch Sampling. Since each dataset uses its own N-point template, all samples within a dataset share the same number of queried landmarks. For each training iteration, we randomly select (without replacement) one dataset and sample a batch (equal to the batch size) from it. This ensures consistent tensor shapes for landmark queries and avoids the need for jagged arrays.

A.13 Explanation of Slight Performance Drop on WFLW68

We address this issue in L282–290 of the manuscript and expand on it here. As noted in L219–220, the WFLW dataset presents a wide range of challenges, including extreme poses, expressions, and occlusions. In our unified training setup, we apply dataset-level oversampling to maintain a balanced exposure across all datasets. However, because other datasets often contain less challenging samples, the model's exposure to difficult WFLW-specific cases is reduced. This can explain the slight performance drop on WFLW68. Importantly, while we observe a decrease in NME on the 68-point version of WFLW, we also observe a performance gain on the full 98-point format. This suggests that the model benefits from the additional diverse data, especially in handling the extra 30 facial points. In other words, the gain in the 30 additional landmarks outweighs the loss in the common 68, indicating that our method generalizes well overall when exposed to a wider variety of N-point formats.

A.14 Broader Impact

Our proposed Unified Dynamic Face Landmark Detection (FLD) framework establishes a foundation for adapting FLD systems to downstream applications. While we do not aim to set new benchmark records, our framework introduces fused dataset training without requiring additional dataset information, thereby enabling effective dynamic landmark prediction. Below, we outline potential positive and negative impacts:

Positive Impacts.

1. Accessibility and Efficiency: By unifying multiple landmark datasets into a single framework, the method reduces the need to train separate models for different datasets. This can lower computational costs and facilitate more accessible deployment of FLD systems, particularly in resource-constrained settings.
2. Scalability and Adaptability: The ability to handle variable landmark layouts allows for more adaptable systems that can be tailored for specific applications, such as facial expression analysis (select face parts only), face direction estimation, FLD stabilization in videos, and medical diagnostics involving facial structures.
3. Potential for Improved Fairness: Training a single model on diverse datasets may mitigate biases that arise from models trained solely on specific datasets, potentially leading to more robust performance across varied demographic and environmental conditions.

4. Cross-Dataset Learning: The FPALP framework can promote cross-dataset learning, encouraging researchers to leverage underutilized datasets and discover new patterns in facial landmark configurations.

Negative Impacts:

1. Dataset Bias Amplification: If the unified dataset disproportionately represents certain demographics or facial structures, the model could inadvertently reinforce existing biases, leading to inaccurate predictions or unfair outcomes.

2. Dependence on Dataset Quality: The proposed approach relies on the accuracy and consistency of landmark annotations. If certain datasets contain noisy or imprecise labels, the model's predictions may propagate these errors, potentially compromising its generalizability and reliability.

A.15 Face Template Alignment Statistics

As detailed in Sec. 3 of the main paper, the first step in the formulation of Face Part-Anchored Landmark Positions (FPALPs) is the synthesis of the unified face template T_{U} through an alignment of the individual face templates of the considered datasets. The unified template is constructed algorithmically, via clustering across datasets. In our implementation, we use the mid-density dataset to set the initial clustering centers, restrict drastic center movement to avoid drift, and flag landmarks that do not fit existing clusters. When newly introduced landmarks yield large member distances, it is treated as a new cluster center / new face part rather than being forcibly merged.

We specified that the alignment of the face templates of AFLW19, 300W, and WFLW, resulted in tight proximal clusters having an average intra-cluster distance of 2.22 pixels averaged over all face parts. In Table 9 we expand this statistic by showing the per-face part mean intra-cluster distances of landmark clusters having at least two landmark members.

We theorize that these intra-cluster distances quantifies a blend of (1) semantic positioning inconsistency across multiple poses, arising from differences in how datasets define contour trajectories under foreshortening, profile rotations, occlusions and different facial expressions, and (2) subjective annotation noise at the annotator level. FPALPs operate under this noisy supervision yet still learn a stable "average" contour representation. A promising direction for future work is to explicitly model and decompose this noise, e.g., by estimating semantic curve variability separately from annotator-level deviations and training the model to minimize the former while remaining robust to the latter.

Table 9: Mean intra-cluster distance (in pixels) for the landmark clusters per face part during the alignment of the face templates of the AFLW19, 300W, and WFLW datasets, into a unified face template. A clean alignment is observed with the minimum, maximum, and mean values of the mean intra-cluster distance taken across the face parts as 1.51, 3.82, and 2.22 pixels respectively.

Face Part	Mean Intra-Cluster Distance
face contour	3.82
left eyebrow	2.11
right eyebrow	2.24
nose bridge	2.30
nose boundary	1.51
left eye	1.59
right eye	1.52
outer lip	1.82
inner lip	3.27
left pupil	2.12
right pupil	2.07

A.16 Landmark to FPALP Mapping

After we attain the unified face template T_{U}, we assign each landmark to one or more user-defined face parts and calculate its Face Part-Anchored Landmark Positions. We tabulate the result of this assignment for the AFLW, COFW, 300W, and 300W datasets in Tab. Table 10-13 respectively.

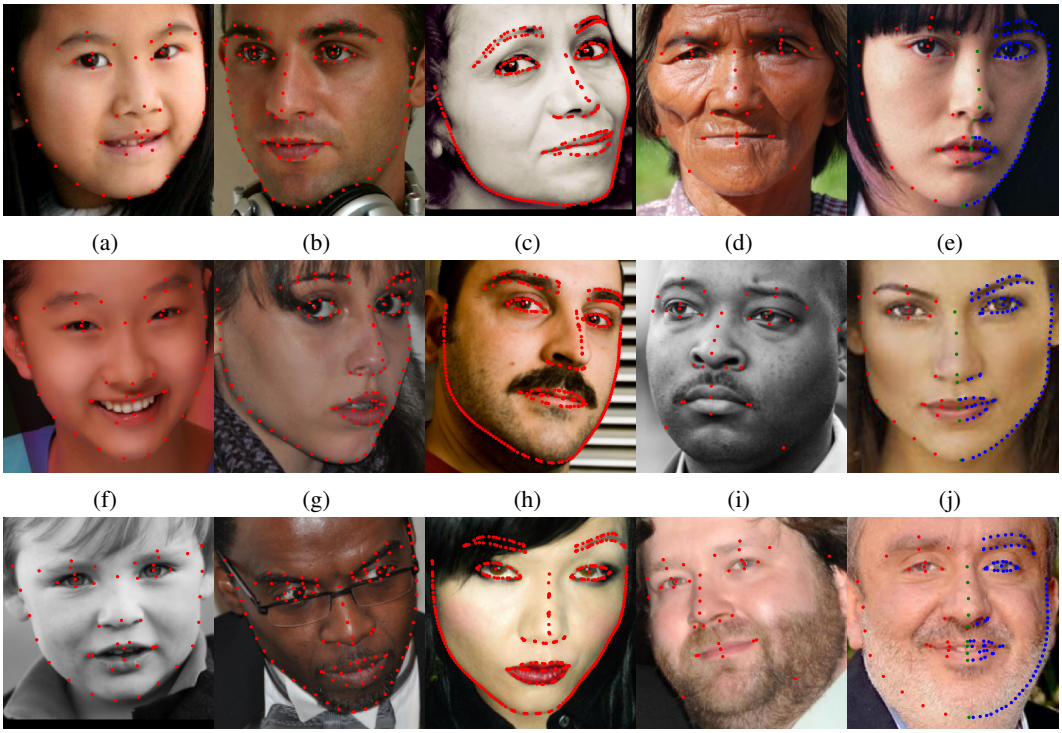


Figure 7: An illustration of the dynamic landmark prediction capability of our system. All images are selected from the WFLW test set which implements the 98-point face template. We anchor the landmarks to the following face parts: left and right eyes, eyebrows, and pupils, inner and outer lips, face contour, nose bridge and boundary. For (e), we split the face contour, lips, and nose boundary into left, center, and right sub-parts. Landmark predictions per face part are depicted using (a)(f)(k) a granularity multiplier of 0.5, (b)(g)(l) a granularity multiplier of 1, (c)(h)(m) a granularity multiplier of 4, (d)(i)(n) 4 landmarks per face part, and (e)(j)(o) a granularity multiplier of 0.5, 1, and 2 for left, center, and right sub-parts whose landmarks are color coded as red, green, and blue respectively.

A.17 Showcasing Dynamic Landmark Prediction: Additional Visualizations

As in Figure 6 within Sec. 4.3 of our main paper, we qualitatively assess the output of our Generalized Dynamic Face Landmark Detection system by depicting a variety of dynamic landmark prediction configurations on images from the WFLW test set in Figure 7. Further, in order to depict the robustness of our system, we visualize our landmark prediction configurations on the challenging cases of occlusion in Figure 8 and extreme poses in Figure 9 from the WFLW test set. From the visualizations, we observe that our framework is able to successfully reason and predict the most likely positions for the landmarks despite the (partial and complete) occlusion of face parts, atypical facial expressions, and extreme poses.

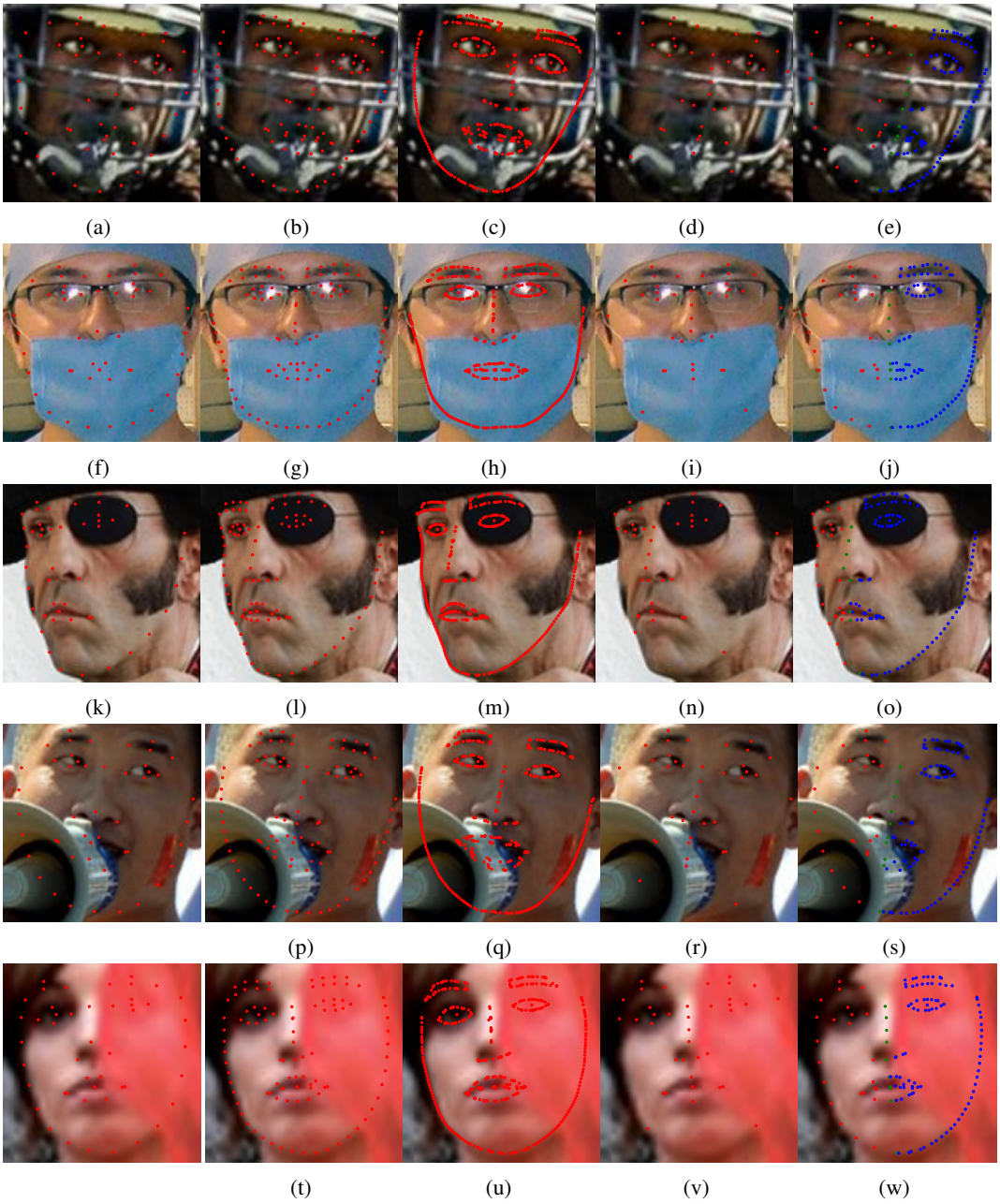


Figure 8: An illustration of the dynamic landmark prediction capability of our system on challenging occlusion cases from the WFLW test set. We anchor the landmarks to the following face parts: left and right eyes, eyebrows, and pupils, inner and outer lips, face contour, nose bridge and boundary. For (e), we split the face contour, lips, and nose boundary into left, center, and right sub-parts. Landmark predictions per face part are depicted using (a)(f)(k) a granularity multiplier of 0.5, (b)(g)(l) a granularity multiplier of 1, (c)(h)(m) a granularity multiplier of 4, (d)(i)(n) 4 landmarks per face part, and (e)(j)(o) a granularity multiplier of 0.5, 1, and 2 for left, center, and right sub-parts whose landmarks are color coded as red, green, and blue respectively.

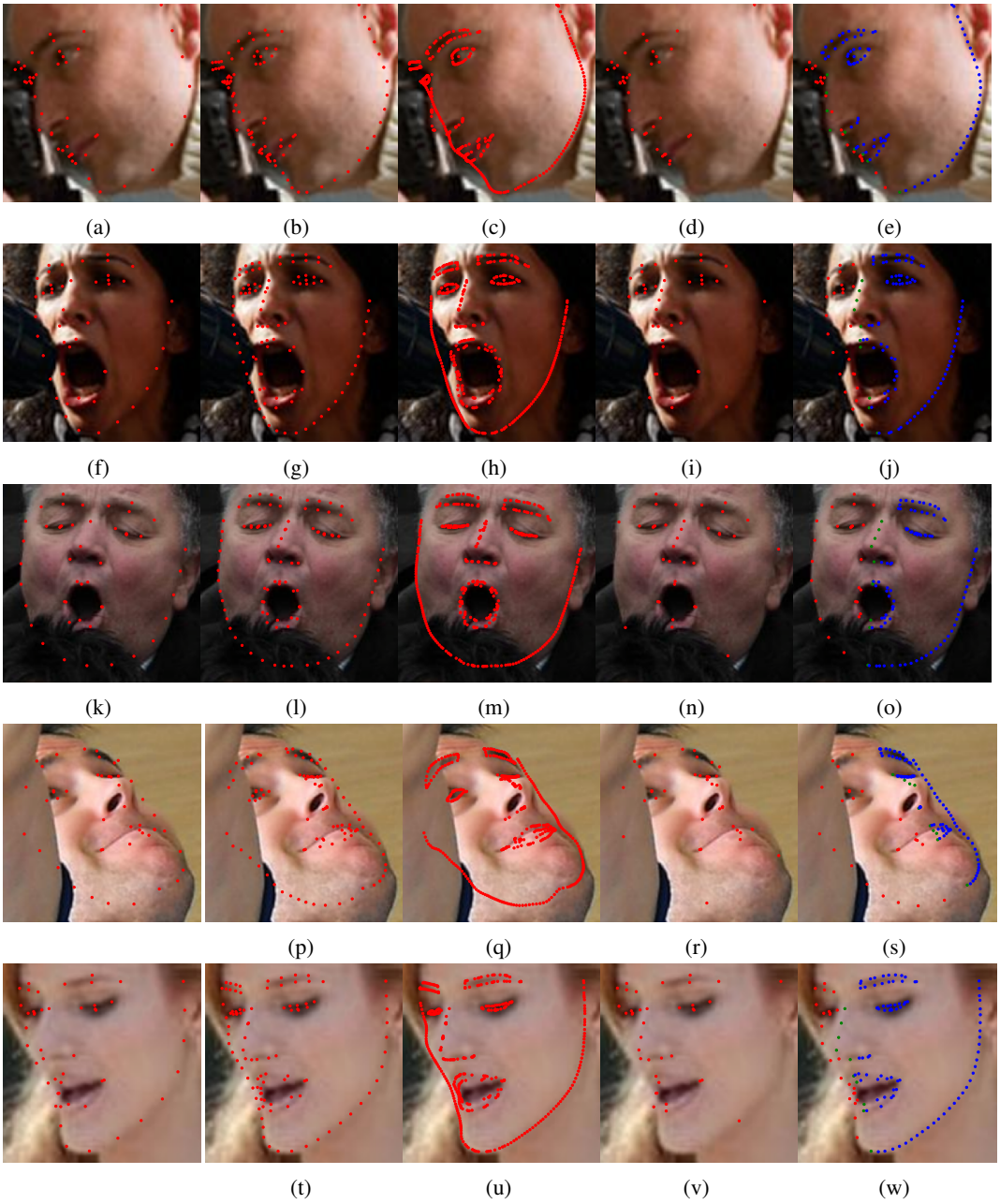


Figure 9: An illustration of the dynamic landmark prediction capability of our system on challenging extreme pose cases from the WFLW test set. We anchor the landmarks to the following face parts: left and right eyes, eyebrows, and pupils, inner and outer lips, face contour, nose bridge and boundary. For (e), we split the face contour, lips, and nose boundary into left, center, and right sub-parts. Landmark predictions per face part are depicted using (a)(f)(k) a granularity multiplier of 0.5, (b)(g)(l) a granularity multiplier of 1, (c)(h)(m) a granularity multiplier of 4, (d)(i)(n) 4 landmarks per face part, and (e)(j)(o) a granularity multiplier of 0.5, 1, and 2 for left, center, and right sub-parts whose landmarks are color coded as red, green, and blue respectively.

Table 10: Mapping of AFLW's 19 landmarks to their face parts and Face Part-Anchored Landmark Positions (FPALPs).

Landmark ID	Face Part	FPALP
1	Face Contour	16/32
2	Left Eyebrow	0/9 or 0/2
3	Left Eyebrow	4.5/9 or 1/2
4	Right Eyebrow	0/9 or 0/2
5	Right Eyebrow	4.5/9 or 1/2
6	Middle of Left Eyebrow	0/1
7	Middle of Right Eyebrow	0/1
8	Nose Bridge	3/3
9	Nose Boundary	0/6
10	Nose Boundary	6/6
11	Left Eye	0/6
12	Left Eye	3/6 or 1/2
13	Right Eye	0/6
14	Right Eye	3/6 or 1/2
15	Outer Lip	0/12
16	Outer Lip	6/12
17	Middle of Mouth	0/1
18	Left Eye Pupil	0/1
19	Right Eye Pupil	0/1

Table 11: Mapping of COFW's 29 landmarks to their face parts and Face Part-Anchored Landmark Positions (FPALPs).

Landmark ID	Face Part	FPALP
1	Face Contour	16/32
2	Left Eyebrow	0/9
3	Left Eyebrow	2/9
4	Left Eyebrow	1/2
5	Left Eyebrow	7/9
6	Right Eyebrow	0/9
7	Right Eyebrow	2/9
8	Right Eyebrow	1/2
9	Right Eyebrow	7/9
10	Nose Bridge	3/3
11	Nose Boundary	0/6
12	Nose Boundary	3/6
13	Nose Boundary	6/6
14	Left Eye	0/8
15	Left Eye	2/8
16	Left Eye	4/8
17	Left Eye	6/8
18	Right Eye	0/8
19	Right Eye	2/8
20	Right Eye	4/8
21	Right Eye	6/8
22	Outer Lip	0/12
23	Outer Lip	3/12
24	Outer Lip	6/12
25	Outer Lip	9/12
26	Inner Lip	2/8
27	Inner Lip	6/8
28	Left Eye Pupil	0/1
29	Right Eye Pupil	0/1

Table 12: Mapping of 300W's 68 landmarks to their face parts and Face Part-Anchored Landmark Positions (FPALPs).

Landmark ID	Face Part	FPALP	Landmark ID	Face Part	FPALP
1	Face Contour	0/32	35	Nose Boundary	4/6
2	Face Contour	2/32	36	Nose Boundary	5/6
3	Face Contour	4/32	37	Left Eye	0/6
4	Face Contour	6/32	38	Left Eye	1/6
5	Face Contour	8/32	39	Left Eye	2/6
6	Face Contour	10/32	40	Left Eye	3/6
7	Face Contour	12/32	41	Left Eye	4/6
8	Face Contour	14/32	42	Left Eye	5/6
9	Face Contour	16/32	43	Right Eye	0/6
10	Face Contour	18/32	44	Right Eye	1/6
11	Face Contour	20/32	45	Right Eye	2/6
12	Face Contour	22/32	46	Right Eye	3/6
13	Face Contour	24/32	47	Right Eye	4/6
14	Face Contour	26/32	48	Right Eye	5/6
15	Face Contour	28/32	49	Outer Lip	0/12
16	Face Contour	30/32	50	Outer Lip	1/12
17	Face Contour	32/32	51	Outer Lip	2/12
18	Left Eyebrow	0/9	52	Outer Lip	3/12
19	Left Eyebrow	1/9	53	Outer Lip	4/12
20	Left Eyebrow	2/9	54	Outer Lip	5/12
21	Left Eyebrow	3/9	55	Outer Lip	6/12
22	Left Eyebrow	4/9	56	Outer Lip	7/12
23	Right Eyebrow	0/9	57	Outer Lip	8/12
24	Right Eyebrow	1/9	58	Outer Lip	9/12
25	Right Eyebrow	2/9	59	Outer Lip	10/12
26	Right Eyebrow	3/9	60	Outer Lip	11/12
27	Right Eyebrow	4/9	61	Inner Lip	0/8
28	Nose Bridge	0/3	62	Inner Lip	1/8
29	Nose Bridge	1/3	63	Inner Lip	2/8
30	Nose Bridge	2/3	64	Inner Lip	3/8
31	Nose Bridge	3/3	65	Inner Lip	4/8
32	Nose Boundary	1/6	66	Inner Lip	5/8
33	Nose Boundary	2/6	67	Inner Lip	6/8
34	Nose Boundary	3/6	68	Inner Lip	7/8

Table 13: Mapping of WFLW's 98 landmarks to their face parts and Face Part-Anchored Landmark Positions (FPALPs).

Landmark ID	Face Part	FPALP	Landmark ID	Face Part	FPALP
1	Face Contour	0/32	50	Right Eyebrow	7/9
2	Face Contour	1/32	51	Right Eyebrow	8/9
3	Face Contour	2/32	52	Nose Bridge	0/3
4	Face Contour	3/32	53	Nose Bridge	1/3
5	Face Contour	4/32	54	Nose Bridge	2/3
6	Face Contour	5/32	55	Nose Bridge	3/3
7	Face Contour	6/32	56	Nose Boundary	1/6
8	Face Contour	7/32	57	Nose Boundary	2/6
9	Face Contour	8/32	58	Nose Boundary	3/6
10	Face Contour	9/32	59	Nose Boundary	4/6
11	Face Contour	10/32	60	Nose Boundary	5/6
12	Face Contour	11/32	61	Left Eye	0/8
13	Face Contour	12/32	62	Left Eye	1/8
14	Face Contour	13/32	63	Left Eye	2/8
15	Face Contour	14/32	64	Left Eye	3/8
16	Face Contour	15/32	65	Left Eye	4/8
17	Face Contour	16/32	66	Left Eye	5/8
18	Face Contour	17/32	67	Left Eye	6/8
19	Face Contour	18/32	68	Left Eye	7/8
20	Face Contour	19/32	69	Right Eye	0/8
21	Face Contour	20/32	70	Right Eye	1/8
22	Face Contour	21/32	71	Right Eye	2/8
23	Face Contour	22/32	72	Right Eye	3/8
24	Face Contour	23/32	73	Right Eye	4/8
25	Face Contour	24/32	74	Right Eye	5/8
26	Face Contour	25/32	75	Right Eye	6/8
27	Face Contour	26/32	76	Right Eye	7/8
28	Face Contour	27/32	77	Outer Lip	0/12
29	Face Contour	28/32	78	Outer Lip	1/12
30	Face Contour	29/32	79	Outer Lip	2/12
31	Face Contour	30/32	80	Outer Lip	3/12
32	Face Contour	31/32	81	Outer Lip	4/12
33	Face Contour	32/32	82	Outer Lip	5/12
34	Left Eyebrow	0/9	83	Outer Lip	6/12
35	Left Eyebrow	1/9	84	Outer Lip	7/12
36	Left Eyebrow	2/9	85	Outer Lip	8/12
37	Left Eyebrow	3/9	86	Outer Lip	9/12
38	Left Eyebrow	4/9	87	Outer Lip	10/12
39	Left Eyebrow	5/9	88	Outer Lip	11/12
40	Left Eyebrow	6/9	89	Inner Lip	0/8
41	Left Eyebrow	7/9	90	Inner Lip	1/8
42	Left Eyebrow	8/9	91	Inner Lip	2/8
43	Right Eyebrow	0/9	92	Inner Lip	3/8
44	Right Eyebrow	1/9	93	Inner Lip	4/8
45	Right Eyebrow	2/9	94	Inner Lip	5/8
46	Right Eyebrow	3/9	95	Inner Lip	6/8
47	Right Eyebrow	4/9	96	Inner Lip	7/8
48	Right Eyebrow	5/9	97	Left Eye Pupil	0/1
49	Right Eyebrow	6/9	98	Right Eye Pupil	0/1